Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Period: \_\_\_\_\_\_

**Activity: Statistical Analysis of Molecular Sequence Data**

Ms. OK, AP Biology, 2014-2015

**Background Information:** Below, you are given the % difference in DNA base sequences between gorillas vs. humans and chimpanzees vs. humans. You will calculate the mean, standard deviation, and standard error of the mean (SEM) for each set of data. Finally, you will graph the means with error bars displaying the 95% confidence limit for each set of data. This will allow you to determine whether there is a statistically significant difference between the % difference in DNA base sequences for these two sets of data.

*Note: This data is TOTALLY made up and inaccurate… but the goal here is to learn statistical analysis of data!!!*

|  |  |  |
| --- | --- | --- |
| **Gene** | **% Difference between Gorilla and Human DNA Base Sequences** | **% Difference between Chimp and Human DNA Base Sequences** |
| 1 | 5 | 4 |
| 2 | 10 | 5 |
| 3 | 7 | 2 |
| 4 | 6 | 4 |
| 5 | 8 | 3 |

1. Use the data in the chart above to calculate the **mean** for each set of data.

***Why use this formula?***

Use the mean formula when you want to calculate the average of a set of values (data points). When you calculate the mean for two sets of values, this allows you to easily compare the sets of values

***Formula***



*Note: I agree with you… this formula makes NO SENSE! The easiest way to calculate mean is simply by adding the data points together and dividing by the number of data points (n).*

***Additional Information from the Formula Sheet***



Record your calculated means in the table given below.

|  |  |  |
| --- | --- | --- |
|  | **% Difference between Gorilla and Human DNA Base Sequences** | **% Difference between Chimp and Human DNA Base Sequences** |
| (Mean) |  |  |

2. Use the data in the chart on the previous page to calculate the **standard deviation** for each set of data.

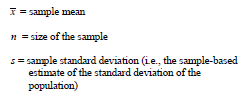
***Why use this formula?***

Use standard deviation formula to determine the amount by which your values (data points) typically differ from the mean value. In other words, the standard deviation determines the amount of variation in your data.

***Formula***



***Additional Information from the Formula Sheet***



Record your work in row 1 of the table below, and record your calculated standard deviations in row 2 of the table given below.

|  |  |  |
| --- | --- | --- |
|  | **% Difference between Gorilla and Human DNA Base Sequences** | **% Difference between Chimp and Human DNA Base Sequences** |
| Work |  |  |
| s (standard deviation) |  |  |

3. Use the standard deviations you calculated to find the **standard error of the mean (SEM)** for each set of data.

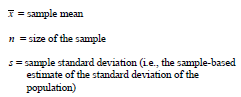
***Why use this formula?***

Use the standard error formula to determine the precision of the mean value. In other words, we are determining how confident we are in our mean value by considering both the standard deviation (s) and the number of data points (n). Typically, when we have more data points, we can be more confident in our data (i.e. a lower standard error).

***Formula***



***Additional Information from the Formula Sheet***



Record your calculated SEM values in the table given below.

|  |  |  |
| --- | --- | --- |
|  | **% Difference between Gorilla and Human DNA Base Sequences** | **% Difference between Chimp and Human DNA Base Sequences** |
| (SEM) |  |  |

4. Next you will calculate the **95% confidence limit.**

***Formula***

There is no formula given on the test, but if you are asked to use the 95% confidence limit to draw error bars on a graph, they will tell you that the 95% confidence limit includes all values within the following range…

95% Cl = Mean ± 2SEM

***Why use this formula?***

The (mean + 2SEM) gives you the **error bar upper limit** (the top of the error bar). The (mean – 2SEM) gives you the **error bar lower limit** (the bottom of the error bar). The error bar provides a range of values around the population mean. If we sampled a much larger number of data points (in this case gene sequences), we would be 95% confident that the mean of this larger sample was within the error bar range.

Record your calculated error bar upper limits in row 1 of the table below, and record your calculated error bar upper limits in row 2 of the table below.

|  |  |  |
| --- | --- | --- |
|  | **% Difference between Gorilla and Human DNA Base Sequences** | **% Difference between Chimp and Human DNA Base Sequences** |
| Error Bar Upper Limit |  |  |
| Error Bar Lower Limit |  |  |

5. In the grid given below, create a bar graph showing the mean % difference in DNA base sequences for gorilla vs. human and chimp vs. human. (You should be graphing only two bars.) For each set of data, please graph an error bar using the error bar upper limit and error bar lower limit. Make sure to include the following elements of a proper scientific graph.

-A title that includes information about the x and y axis

-X and Y axis labels (with units where applicable)

-An appropriate scale on your X and Y axes

6. Is there overlap in the error bars for the two data sets? What does this mean?

7. Based on the means for each set of data, which species is more closely related to humans—gorillas or chimpanzees? Does the statistical analysis you did allow you to make this claim confidently? Why or why not?